# Assessing Reliability, Heritability and General Cognitive Ability in a Battery of Cognitive Tasks for Laboratory Mice

Galsworthy, Michael J ; Paya-Cano, Jose L ; Liu, Lin ; Monleón, Santiago ; Gregoryan, Gregory ; Fernandes, Cathy ; Schalkwyk, Leonard C ; Plomin, Robert

Abstract: This report includes the first sibling study of mouse behavior, and presents evidence for a heritable general cognitive ability (g) factor influencing cognitive batteries. Data from a population of male and female outbred mice (n = 84), and a replication study of male sibling pairs (n = 167) are reported. Arenas employed were the T-maze, the Morris water maze, the puzzle box, the Hebb-Williams maze, object exploration, a water plus-maze, and a second food-puzzle arena. The results show a factor structure consistent with the presence of g in mice. Employing one score per arena, this factor accounts for 41% of the variance in the first study (or 36% after sex regression) and 23% in the second, where this factor also showed sibling correlations of 0.17-0.21, which translates into an upper-limit heritability estimate of around 40%. Reliabilities of many tasks are low and consequently set an even lower ceiling for inter-arena or sibling correlations. Nevertheless, the factor structure is seen to remain fairly robust across permutations of the battery composition and the current findings fit well with other recent studies

DOI: https://doi.org/10.1007/s10519-005-3423-9

*Behavior Genetics, Vol. 35, No. 5, September 2005 (© 2005)*
DOI: 10.1007/s10519-005-3423-9

# Assessing Reliability, Heritability and General Cognitive Ability in a Battery of Cognitive Tasks for Laboratory Mice

**Michael J. Galsworthy,[1,2,5] Jose L. Paya-Cano,[1] Lin Liu,[1] Santiago Monleón,[3] Gregory Gregoryan,[4] Cathy Fernandes,[1] Leonard C. Schalkwyk,[1] and Robert Plomin[1]**

*Received 19 Aug. 2004—Final 09 Feb. 2005*

This report includes the first sibling study of mouse behavior, and presents evidence for a heritable general cognitive ability (*g*) factor influencing cognitive batteries. Data from a population of male and female outbred mice ($n = 84$), and a replication study of male sibling pairs ($n = 167$) are reported. Arenas employed were the T-maze, the Morris water maze, the puzzle box, the Hebb–Williams maze, object exploration, a water plus-maze, and a second food-puzzle arena. The results show a factor structure consistent with the presence of g in mice. Employing one score per arena, this factor accounts for 41% of the variance in the first study (or 36% after sex regression) and 23% in the second, where this factor also showed sibling correlations of 0.17–0.21, which translates into an upper-limit heritability estimate of around 40%. Reliabilities of many tasks are low and consequently set an even lower ceiling for inter-arena or sibling correlations. Nevertheless, the factor structure is seen to remain fairly robust across permutations of the battery composition and the current findings fit well with other recent studies.

**KEY WORDS:** Factor analysis; *g*; general cognitive ability; heritability; HS mice; individual differences; siblings.

## INTRODUCTION

In humans, it has long been known that an individual's performance on one cognitive task is reasonably predictive of performance other cognitive tasks. This was first documented by Spearman (1904), who accounted for the phenomenon by coining *g*, short for "general cognitive ability", as an underlying cognitive trait which is tapped into by all cognitive tasks. One century on, Spearman's g has strengthened and evolved as hundreds of psychological and behavioral genetic investigations have validated the concept and shown g to be one of the most stable and heritable of all human behavioral traits (Brody, 1992; Deary, 2000; Mackintosh, 1998; Plomin *et al*., 2001). Not only is *g* critical for the investigation of mental retardation—manifest as impairments in general functioning—it also impinges on investigations of specific abilities or disabilities (Plomin, 1999).

However, there has been a dearth of parallel research in mice or rats. As a result, there is no adequate rodent model of *g* with which to explore the functional genomics of the phenomenon (Plomin, 2001). Studies employing outbred mice on multiple cognitive tasks are rare and this has hampered adequate psychometric ascertainment of whether such tasks overlap in measurement. This is not only an impediment to understanding the individual differences structure of

[1] Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, King's College London, London, UK.
[2] Division of Neuroanatomy and Behavior, Institute of Anatomy, University of Zurich, Winterthurerstrasse 190, Switzerland CH-8057.
[3] Área de Psicobiología, Facultad de Psicología, Universitat de Valencia, Spain.
[4] Institute of Higher Nervous Activity and Neurophysiology, Russian Academy of Sciences, Moscow, Russia.
[5] To whom correspondence should be addressed at Division of Neuroanatomy and Behavior, Institute of Anatomy, University of Zurich, Winterthurerstrassé 190, CH-8057 Zurich, Switzerland. Tel: ++411 6355 359; Fax: ++411 6355 702; e-mail: m.galsworthy@anatom.unizh.ch.

cognition in mice, but it also means studies employing different cognitive tasks cannot confidently claim relevance to each other on levels of normal genetic or environmental variation.

Until recently, the only study as far as the authors are aware which provides individual differences data in mice across various cognitive tasks is that of Bagg (1920). The experiment was not designed to explore *g*, and only reported one cross-arena correlation. Nevertheless, the manuscript made available raw data for a large subset of the mice, and these data have been reanalyzed here (see Appendix). In summary, latency and error scores for 4 tasks in two arenas are all seen to inter-correlate positively, indicating a common trait determining the quality of performance across all eight measures. This is clarified by an unrotated principal component factor analysis showing that all measures load in the same direction on a first factor accounting for 61% of the task variance. We choose to employ this as a preliminary indicator of *g*, as uniform direction of loading confirms that all measures covaried in a manner consistent with a *g* hypothesis. However, it must be admitted that this does not confirm the variance to be exclusively cognitive. In this example of the Bagg data, it may be argued that the same motivation in all tasks (namely desire to gain access to a community area) represents another factor promoting consistency of performance.

More recently, Locurto and Scanlon (1998) assessed C57BL/6 × DBA/2J F2 and CD-1 populations of mice on a battery of spatial tasks under water-escape motivation. All cognitive tasks were positively inter-related with a general factor accounting for between 28% and 61% of the task variance. Yet as with the Bagg data, the use of the same motivational demands throughout the battery meant that any general factor extracted from the battery may not have been exclusively cognitive. Three studies since then have tackled directly this issue of motivational confound. Galsworthy *et al.* (2002) employed a battery of diverse cognitive tasks that spanned wet and dry arenas under varying motivations. All tasks were seen to load positively on a general factor accounting for approximately 30% of the variance. Locurto *et al.* (2003) ran a similar battery. In their principal component factor analysis, which included three control measures, not all cognitive tasks loaded in the same direction, and so three rotated factors were presented as explaining the cognitive variance. Matzel *et al.* (2003), however, found results more comparable to Galsworthy *et al.* (2002) for their diverse battery assessing learning rates. All measures loaded in the same direction on a first unrotated factor accounting for 38% of the battery variance.

Table I summarizes the inter-arena correlations for all relevant cognitive studies in mice. It can be seen that although there is a strong preponderance of positive inter-arena correlations, the mean correlation magnitudes are low, especially for batteries spanning multiple motivations. Note also that no significant ($p < 0.05$) negative inter-arena correlations have been found in any of these studies.

**Table I.** Summary of Inter-arena Correlations for Cognitive Tasks in Mice

| | | | No. of correlations | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Negative | | Positive | | | |
| Study | *n* | Motivation | $p < 0.05$ | n.s. | n.s. | $p < 0.05$ | Mean *r* | *g* |
| Bagg (1920)[a] | 71 | Within | 0 | 0 | 2 | 10 | 0.40 | 61% |
| Locurto and Scanlon (1998)[b] | 34 | Within | 0 | 0 | 5 | 16 | 0.36 | 28–61% |
| Locurto and Scanlon (1998)[c] | 41 | Within | 0 | 0 | 4 | 17 | 0.37 | 37–55% |
| Galsworthy *et al.* (2002) | 40 | Cross | 0 | 1 | 16 | 6 | 0.20 | 28–31% |
| Locurto *et al.* (2003)[d] | 60 | Cross | 0 | 11 | 32 | 14 | 0.12 | n/a |
| Matzel *et al.* (2003) | 56 | Cross | 0 | 0 | 8 | 2 | 0.22 | 38% |

[a]Correlations and conclusions concerning *g* first presented here (see Apendix), derived from data published in Bagg (1920).
[b]$F_2$ population, all latency and error correlations.
[c]CD-1 population, all latency and error correlations.
[d]All error, latency, errorless trial and aggregate score correlations.
"Within" indicates a cross-arena study but employing a uniform motivational drive, "Cross" indicates studies employing varied motivational demands.
"$p < 0.05$" indicates significant as reported in the study, and "n.s." indicated non-significant as reported in the study.
"g" displays the magnitudes of first factors for studies where a common factor was concluded.

The aim of the research reported here was to test the hypothesis of a *g* factor in mice, explore the reliabilities of our cognitive tasks, and estimate familial (genetic and shared environmental) contributions to these by use of sibling correlations. This work represents part of a larger study to develop cognitive tasks suitable for genetic association (quantitative trait locus, QTL) and functional genomic exploration of natural cognitive variation in laboratory mice.

## MATERIALS AND METHODS

### Study Design

Two aspects of the study methodology are described here in overview before giving the specific details of mice and tasks. The first of these is the study structure and the second is the design of the cognitive battery.

With regard to overall structure; four batches of mice were employed. Study 1 consists of the first two batches in which unrelated mixed-sex mice were used. This included both pigmented and albino mice. Study 2 consists of a further two batches employing male-only sibling pairs. Albinos were excluded (see Section "Descriptive Statistics"). In some cases procedures differed between batches within a study. Therefore, when adding the two component batches of a study, scores were standardized within batch before the datasets were added. In Study 2, the sample size was increased substantially so as to be able to detect as significant the low to moderate correlations being found (to detect a correlation of 0.20 at the 5% level with a two-tailed test and 80% power, a sample size of 194 is needed; see Cohen, 1988). Data from the first batch have been presented before (Galsworthy *et al.*, 2002), and some of these data are subsumed in the Study 1 dataset reported here.

With regard to the design of the cognitive battery; tasks were sought out that were presumed to tap higher-level cognitive functioning such as working memory, spatial navigation, complex object manipulation and problem-solving. Considerations such as low stress and lack of cost in terms of money and time also influenced task choice. Therefore long operant schedules or shock-based measures such as fear conditioning were not included. For Study 2, the original set of chosen tasks (namely: spontaneous alternation, two puzzle-box tasks, the Morris maze and Hebb–Williams maze) was expanded to include an object exploration task (as this has been argued to be closely associated with cognitive task performance; see "Discussion"), a water plus maze (to compare with the Morris maze), and the syringe puzzle (an additional object manipulation task). Finally, the tasks were chosen such that there was an overall balance in the battery between well-established tasks and newly developed tasks, between water-based tasks and land-based tasks, between spatial navigation and non-spatial-navigation tasks, between punishing errors and encouraging exploration, and covering a range of provoked motivational drives. Thus, it is argued that the continuing thread of the battery was one of cognitive demand (perhaps excepting the object exploration), with a randomization of other factors. The tasks employed in both studies, plus their presumed primary cognitive demands and motivations, are displayed in Table II.

**Table II.** Summary of the Cognitive Battery Design

| Task | Cognitive process | Motivation | Key reference |
|---|---|---|---|
| T-maze | Working memory (spatial) | Exploration | Gerlai (1998) |
| Burrowing puzzle | Problem solving | Compound[a] | Galsworthy *et al.* (2002) |
| Plug puzzle | Problem solving (manipulation) | Compound[a] | Galsworthy *et al.* (2002) |
| Hebb–Williams maze | "Intelligence", route learning | Water escape[b] | Meunier *et al.* (1986) |
| Morris water maze | Spatial navigation | Water escape | Morris (1984) |
| Water plus maze | Spatial navigation | Water escape | Locurto and Scanlon (1998) |
| Object exploration | Response to novelty | Exploration | Anderson (1993) |
| Syringe puzzle | Problem solving (manipulation) | Food | Galsworthy (2003) |

[a]"Compound" denotes the deliberate use of various motivational drives in parallel: In this case, a strong light/dark difference between the start and goal box, objects to explore and hide in within the goal box, and a small entrance to the goal box.
[b] This water escape is marked as being different from the Morris and water plus mazes, as the mice are not swimming, but wading. Also the temperature is colder, being 15°C instead of 21°C.
*Note:* "manipulation" indicates a spatial component, but this is more akin to human "spatial" object manipulation tasks than to classical animal "spatial" navigation.

**Subjects**

All mice were obtained from the Institute for Behavioral Genetics (IBG) at the University of Colorado at Boulder. IBG HS mice are systematically outbred stock established over 30 years ago from an eight-way cross of C57BL (note: not C57BL/6), BALB/c, RIII, AKR, DBA/2, Is/Bi, A and C3H/2 inbred mouse strains (McClearn *et al.*, 1970). On arrival in the UK, animals were housed individually and maintained in a reversed 12-hour light/dark cycle in an environment controlled for temperature (21 ±2 °C) and humidity. Food (Rat & Mouse No. 1 Maintenance Diet, Special Diet Services, Essex, UK) and water were available *ad libitum*.

In Study 1, 84 heterogeneous stock (HS) mice were used, divided into a first batch of 40 mice (HS generation 64) and a second batch of 44 mice (HS generation 65). Equal numbers of males and females were present in both batches, both pigmented mice and albinos were included. In Study 2, 170 pig-

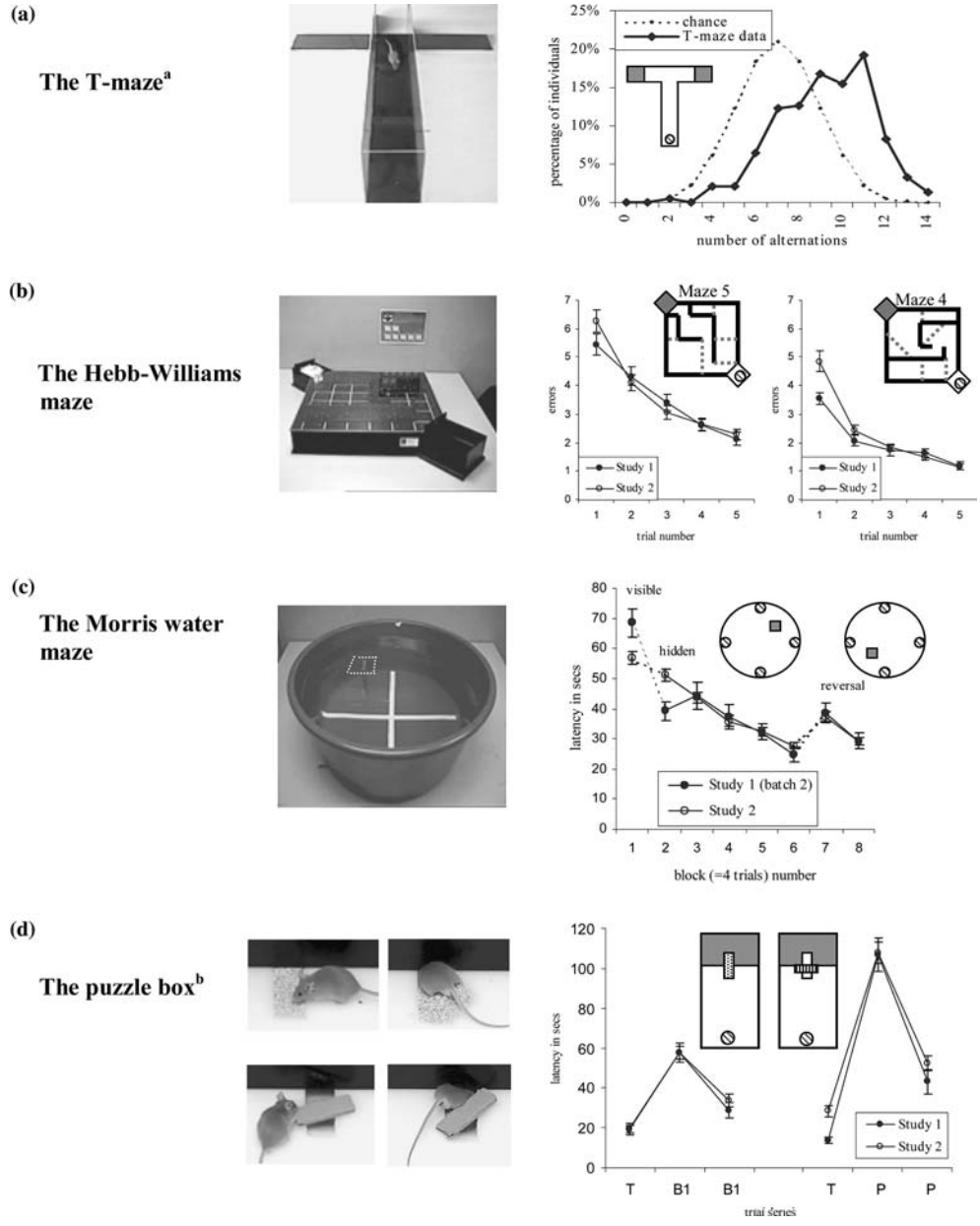


**Fig. 1.** (*Continued*)

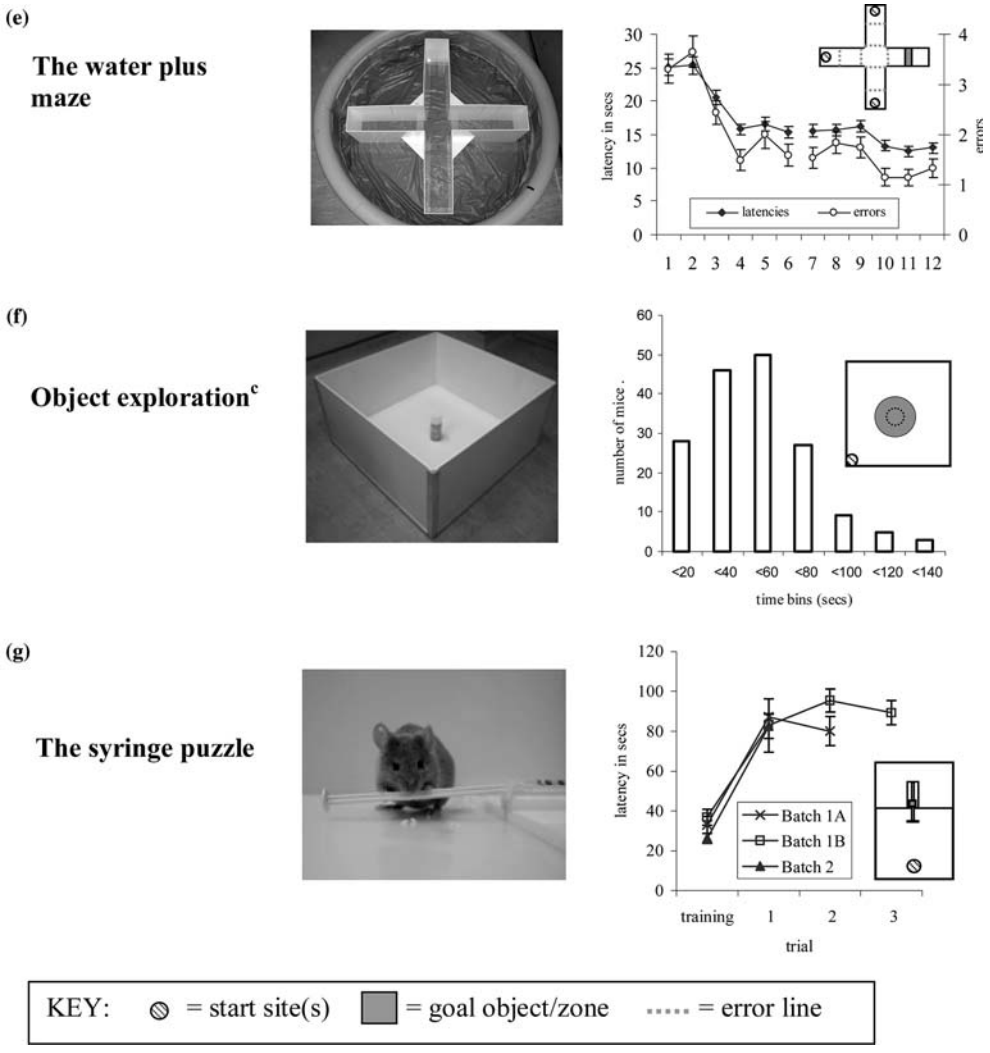


**Fig. 1.** The cognitive arenas employed on both Study 1 and Study 2. (a) The T-maze[a]. (b) The Hebb–Williams maze. (c) The Morris water maze. (d) The puzzle box[b]. (e) The water plus maze. (f) Object exploration[c]. (g) The syringe puzzle. All bars are standard error bars. In the schematics (not to scale) goal objects/zones are colored grey, start points are hatched circles and error zones (Hebb–Williams and water plus maze) are marked by dotted grey lines—as shown in the key at the bottom of the figure. Notes: [a] $n = 245$ mice which completed all 14 trials if the T-maze. [b] "T" = training trial (open underpass), "B1" = first burrowing puzzle where the underpass is filled with sawdust, and "P" = plug puzzle. [c] Data shown in histogram is time spent exploring the novel object.

mented HS mice were used. The group again consisted of two batches, a first batch of 80 males (=40 sibling pairs), followed by a second batch of 90 males (=45 sibling pairs). All mice were from generation 66, with parental combinations changed between batch productions so that the mice in batch 2 were mostly (maternal and paternal) half-siblings of the mice in batch 1.

Two weeks of acclimatization were allowed before cognitive testing, which began when the mice were between 59 and 85 days old for Study 1 and between 95 and 112 days old for Study 2. Two mice from the Study 2 batch 1 died very early in the testing, and one mouse in the Study 2 batch 2 developed a swimming problem in the water plus maze and so was excluded from that task. Cross-maze analyses in Study 2 therefore involve 167 mice with complete data. All procedures carried out on the mice in this study were in compliance with the UK Animal Scientific Procedures Act, 1986 under license from the UK Home Office.

**Testing Arenas**

*T-maze*: This arena is most commonly used for spatial working memory tasks, and has been shown to be sensitive to hippocampal damage (Gerlai, 1998). Two T-mazes were used here: For Study 1, the apparatus employed consists of one longer start arm (36 cm) and two shorter arms (18 cm) forming a T-shape. Opaque lifting doors are located 5 cm into the short arms. The arms are all 6.5 cm wide and 20 cm high. The floor of the maze is black plastic and the walls are clear plastic. For Study 2, a larger T-maze similar to the specification of Gerlai (1998) was employed. It consists of one longer start arm (75 cm) and two shorter arms (31.5 cm), with two sliding opaque doors located 0.2 cm into the short arms. The arms are all 12 cm wide and 20 cm high. The floor of the maze is black plastic and the walls are clear plastic (see Fig. 1a).

*Hebb–Williams maze*: The first notable attempt to develop a standardized set of tasks to study animal "intelligence" produced closed-field and elevated-pathway route-finding mazes for rats (Hebb and Williams, 1946). The closed-field test was further developed and standardized by Rabinovich and Rosvold (1951), with a smaller version for mice then being developed by Meunier *et al*. (1986). A swimming version of this maze for mice was shown to correlate with the Morris water maze and other water-motivated spatial tasks (Locurto and Scanlon, 1998). The model employed here follows the dimensions described by Meunier *et al*. (1986): The maze is made of black plastic 60 cm × 60 cm × 10 cm high, with a start box and a goal box (both 14 cm wide × 9 cm long) at diagonally opposite corners. The maze contains cold water at a wading depth (15 °C, 3.5 cm high), but the goal box was stocked with fresh dry tissues and was covered. Differing arrangements of barriers are fixed to a clear plastic ceiling to produce the various maze designs (selected from Rabinovitch and Rosvold, 1951; see Fig. 1a).

*Morris water maze*: The Morris water maze was first described over 20 years ago as a place navigation task for rats, and most notably showed sensitivity to hippocampal damage (Morris, 1984). The task is now widely used in many different sizes and forms for mice and, despite cautions (Lipp and Wolfer, 1998), remains the most frequently used paradigm to assess "spatial learning", "hippocampal function" or "cognition" in rodents (Ashe, 2001; D'Hooge and De Deyn, 2001). For our version of this task, a mid-blue circular plastic molded tub of 60 cm diameter and 28 cm height was employed. The small diameter was to ensure rapid learning and decreased thigmotaxis (wall-hugging). The maze was filled with water (which was not colored as the underwater platform is not visible from the level of the water surface) 22 cm deep. The platform was a square 6 cm × 6 cm, elevated 1 cm above the water in the visible platform task and submerged 1 cm under the water level in the hidden platform and reversal tasks. Visual cues available were small markings with tape at "north" and "south" points on the wall within the maze, plus posters on the walls outside the arena (see Fig. 1c).

*Puzzle box*: This apparatus was designed to present mice with a series of quick ethological problem-solving tasks, in order to complement spatial navigation tasks in a cognitive battery and more closely resemble human intelligence tasks (Galsworthy *et al*., 2002; Galsworthy, 2003). The intention follows that of Anderson's (1993) "response-flexibility" tasks for rats and also adapts a "burrowing detour" task from Crinella and Yu's (1995) *g*-battery into the puzzle-box arena. The burrowing detour task in rats has been shown to be sensitive to damage in 41 brain areas (Thompson *et al*., 1989, 1990). A variety of other newly developed tasks are also presented in the puzzle-box arena—and so mice are required to dig, climb, push doors, or manipulate and remove objects in order to gain access to the goal box (Galsworthy, 2003). The apparatus is a box 73 cm long × 28 cm wide × 27.5 cm high, divided by a removable barrier into a small dark goal box (14 cm long) containing sawdust and cardboard shapes and a large brightly-lit (~1000 lux) start box (58 cm long). Access to the goal box is normally via an underpass 4 cm wide, 2 cm deep and 15 cm long. This is the entrance blocked by either sawdust (burrowing puzzle) or a T-shaped cardboard plug (plug puzzle) in the two puzzle tasks employed here (see Fig. 1d).

*Water plus-maze*: This arena is a plus-shaped frame inside a large Morris maze and designed for water navigation tasks (Locurto and Scanlon, 1998). Thus it is similar to the Morris water maze, but has the advantages of reducing the effects of thigmotaxis (see Lipp and Wolfer, 1998) and introducing error counts. Both latencies and errors in this arena have been shown to correlate with performance in the Morris water maze and the swimming version of the Hebb–Williams maze (Locurto and Scanlon, 1998). The frame was 30 cm high and placed in 13 cm deep water. The arms were made from white plastic, and internal dimensions of the arms were 14.5 cm wide

and 45 cm long. A platform submerged 1 cm below the water level platform was placed 20–25 cm down one of the four arms, spanning the width of the arm (see Fig. 1e).

*Object exploration*: The novel object exploration task assesses attention to novelty and exploration of a novel, non-aversive object placed in a familiar environment (Misslin and Ropartz, 1981; van Gaalen and Steckler, 2000). In rats, novel object exploration has been shown to correlate with cognitive tasks (Anderson, 1993). The arena employed is essentially an "open field": a box with white plastic walls and floor of internal dimensions $72 \times 72 \times 33$ cm high. An object is introduced to the middle of this arena during the experiment. Illumination was 200 lux provided by a single lamp in an otherwise darkened room (see Fig. 1f).

*Syringe puzzle arena*: This newly developed task is a manipulation task in which mice must pull out a plunger from a syringe in order to gain access to chocolate (Galsworthy, 2003). Like with the puzzle box tasks (see above), the host arena is designed to run a variety of short manipulation or digging tasks. In this case, the apparatus is a small open box which fits different "floors" on which food-reward puzzles are mounted (Galsworthy, 2003). The arena is made from white plastic of internal dimensions $30 \times 44 \times 29.5$ cm high. The syringe is a standard 1 mL syringe (no needle attached) 8.7 cm long and 0.7 cm in diameter, with the rubber end of the plunger removed so that it moves more freely. This is mounted on a plastic piece $7.5 \times 2 \times 0.3$ cm high, which is fixed in turn to the floor-piece ($30 \times 15.5 \times 0.3$ cm thick) centrally to the side walls, but flush with the front edge that the mouse will approach. This effectively raises the syringe (plunger end) some 0.6 cm high. Illumination was ~20 lux diffuse light (see Fig. 1g).

Arena illumination provided for these arenas was room lighting (ranging from 150 to 300 lux), unless specified otherwise above.

## Procedures for Study 1

Due to the ongoing development of the cognitive battery in the laboratory, some tasks and measures differed between the two batches. All the cognitive tasks that were run in both batches are reported below with any procedural differences noted. The order of tests for batch 1 is as given below. The order for batch 2 was spontaneous alternation, burrowing puzzle, plug puzzle, Hebb–Williams maze, Morris water maze.

*Spontaneous alternation (in T-maze)*: The procedure follows that of Gerlai's continuous alternation (Gerlai, 1998). On the first trial, one short arm door was shut. After the mouse had explored the open arm and returned to the start, 14 trials began. Both arms were opened and as the mouse entered one arm, the door to the other arm was quietly closed. When the mouse returned to the start area, both doors were opened and the next trial began. *Measures*: All 14 trials were run in one session and the measure used was number of alternations (0–14). Mice failing to complete 14 trials within the 30 min time limit were awarded only the number of alternations they had completed to that time.

*Hebb–Williams maze*: Mice must navigate the maze from start box to dry goal box to escape the cold water. In batch 1, a 5 min habituation (dry arena, no barriers) session was given on day 1, followed by practice problems A on day 2 and D on day 3 (4 trials/day). Then mazes 1, 5, 3, 4 and 8 were run, each on a separate day employing 8 trials (see Rabinovitch and Rosvold, 1951 for all maze designs). The time limit to find the goal box was 5 min, after which the mouse was guided. Mice in batch 2 were given no dry-arena habituation session, and practice problem A was replaced by training with no barriers. Then mazes 3, 5, and 4 were run in that order, each on a separate day. Each maze was administered 6 times, with a time limit of 3 minutes *Measures*: A total latency score was taken as the summed latencies across all problem trials and mazes. A similar total was used for error scores (where errors were counted as entering an error zone specified by Rabinovitch and Rosvold, 1951).

*Morris water maze*: Four visible platform trials on day 1 (platform 1 cm above water level) were followed by the hidden platform task (platform 1 cm below water level): In batch 1, 32 hidden platform trials were run (days 2–5: 8 trials/day). In batch 2, 20 hidden platform trials were run (4 trials on day 1 + days 2–3: 8 trials/day). A reversal task was run in batch 2 (see Fig. 1C and Section "Procedure for Study 2" for procedure), but this was not included in the Study 1 score due to equivalent data not being available for batch 1. *Measures*: For all trials there was a 60 seconds limit, after which the mouse was guided. The measure used was the summed latencies to find the platform in the hidden platform trials.

*Burrowing puzzle (in the puzzle box)*: Each day consists of 3 trials run in quick succession (inter-trial interval 30–60 seconds). For batch 1, days 1–2 were training trials only (i.e. with a simple black barrier and

an open underpass). For puzzle days, trial 1 is a repetition of the previous day's trial 3, then a novel challenge is presented on trial 2 and repeated on trial 3. The puzzles were as follows: On day 3, the underpass was filled with sawdust and the solution was to burrow through. On day 4, sawdust 1 cm deep covered the floor at the base of the barrier and mice needed to find the location of the underpass and dig through. On day 5, the underpass was again under the sawdust, but blocked with plastic. However, there was now a window in the barrier 12 cm above the underpass, through which the mice could climb. On day 6, the window was removed and the underpass unblocked, repeating the puzzle of day 4. In batch 2, the procedure was identical, but only days 1–4 were run. *Measure*: The measure used for this task was the summed latencies to enter the goal box over all puzzle trials.

*Plug puzzle (in the puzzle box)*: In the plug puzzle, mice are presented with a T-shaped cardboard "plug" blocking the underpass. The plug weighs 2 g and consists of a 2.5 × 7.5 cm (0.5 cm thick) piece sitting across the top of the burrow attached to (but offset 1 cm forward from) a 3 cm wide × 2 cm long × 1.5 cm deep block that sits loosely in the burrow. The plug must be pulled out in order to expose the entrance. Attempts at lifting or pushing will be unsuccessful. As with the burrowing puzzle, the procedure is always that there are three trials per day in quick succession. For batch 1, mice were food deprived for 20 hour beforehand and chocolate was given in the goal box on day 1 only. On the testing day, there was one refresher trial, followed by two trials where the underpass to the goal box was blocked with the plug. There was no chocolate reinforcement on this day as the place preference was already set. The procedure for batch 2 was identical to batch 1, except that there was no food reinforcement (high goal-box exploration with low food consumption on previous trials indicated that food did not substantially add motivation). Also, there was only the test day and no training day. *Measure*: The measure used is the summed latencies to enter goal box on the two problem trials.

### Procedures for Study 2

Order of tasks and procedures were the same as Study 1 batch 2, but there was some shortening, making for the following differences.

In the *T-maze*, the new larger maze was used to standardize with dimensions described elsewhere (Gerlai, 1998); see Section "Testing Arenas". The procedure was otherwise identical. In the *Hebb–Williams maze*, mice were given 1 training day (4 trials) with no barriers. Three mazes were then used (one per day): Mazes 3, 5, 4 for batch 1 (6 trials/day) and 1, 5, 4 for batch 2 (5 trials/day). The change from maze 3 to maze 1 occurred because the correct solution to maze 3 was to follow the left-hand wall—a strategy that most mice had adopted during training. In the *Morris water maze* on day 4, a reversal task was run (8 trials) where the platform was moved to the opposite quadrant. The reversal task trials summed (minus the first trial) was taken as a separate measure for the analyses. In the *burrowing puzzle*, the procedure for batch 1 was a 3-day task equivalent to days 2, 3, and 4 of Study 1, but with the very first trial employing an open 4 × 4 cm doorway above the burrow (to enlarge the entrance and promote exploration). For batch 2, a 1-day (3 trials) procedure was employed in an attempt to develop a very brief cognitive test: trial 1 was the open 4 × 4 cm doorway, trial 2 was the normal training (open underpass), and on trial 3, the underpass was filled with sawdust. The *plug puzzle* procedure was identical to Study 1 batch 2.

Three more tasks were then added to the battery.

*Water plus maze*: In this task, mice navigate to a submerged platform 20 cm down one arm, starting from the ends of the other three arms in the repeating sequence of left-, right-, opposite-arm starts (2 days; 6 trials/day). An error was awarded for each entry into a wrong arm, and a further error awarded for going beyond 30 cm into a wrong arm. The time limit was 60 seconds after which the mouse was guided. *Measures*: A latency total (all latency scores except trial 1, summed), and an error score (all error scores except trial 1, summed) were taken.

*Object exploration*: This task was adapted from Anderson's (1993) object exploration task for rats. Mice were started in one corner and allowed 2 minutes of exploration before a soft drink can was introduced into the centre of the arena. During the following 3 minutes, the mouse's exploration of the can as defined by direct exploratory contact (sniffing, leaning or climbing/being on top of can) was recorded to serve as the measure of object exploration.

*Syringe puzzle*: Mice were deprived of food approximately 20 hours before testing. Due to initial low success rate with this task on batch 1, three versions of the task were run. *Procedure 1 (batch 1, mice 1–40)*: Day 1, trial 1: habituation (3 minutes) to the arena with no syringe. Mice were placed at the start (see Fig. 1c) and allowed to explore. Trial 2:

training trial in which the plunger of the syringe was pulled 0.5 cm out to expose a small clump (~150 mg) of white chocolate 3.5–4.5 cm down the plunger (from the thumb pad end). Time limit was 90 seconds to begin eating, after which 15 seconds were given before the mouse was removed. Mice were returned to their home cage and standard food was returned to the hopper. Day 2: two puzzle trials were run, which were similar to the training trial, but the plunger was pushed in (the chocolate was 1–2 cm inside the syringe tube). Mice had to pull the plunger out to gain access to the chocolate. Time limits were 90 and 120 seconds respectively for the trials. *Procedure 2 (batch 1, mice 41–80)*: Day 1, trial 1: training trial as before. Trial 2: puzzle trial as before. Day 2: both trials were puzzle trials as before. Time limit = 120 seconds for all trials. *Procedure 3 (batch 2)*: One day (two trials) only: a training trial (time limit 120 seconds) and a puzzle trial (time limit 180 seconds). *Measures*: The latency to begin consuming the chocolate on the puzzle trials (summed).

### Analyses

For measures where latency to complete the task was taken, maximum latencies were scored for individuals who were unable to complete the task within the time limit. As noted earlier, all data were standardized within each batch before batch datasets were added within each study. Standardization converts all scores to a common scale where the mean score is 0 and the standard deviation is 1. This controls for not only procedural differences between batches of a study, but also mean effects of environmental differences between batches, which could artificially inflate correlations. All > 24,000 training and problem trials reported in these two studies, plus associated variables, were maintained in a Microsoft Access 97 database. Microsoft Excel and StatTransfer (Circle Systems, Seattle, WA, USA) were then used to re-organize and transfer selected data into Stata (StataCorp 2003. *Stata Statistical Software: Release 8.0*. College Station, TX, USA) for analysis.

Principal component factor analysis (PCFA) was used rather than principal components analysis (PCA) or principal factors analysis (PFA) due to the more interpretable statistics it yields (see Rencher, 1995). However, note that the results here were replicated qualitatively (direction of loadings and their relative magnitudes) by PCA and PFA. Only the unrotated first factors are considered here as the method is being used to test the hypothesis that all measures load in the same direction on a first principal component. The associated binomial probability of such an occurrence being a chance event is $0.5^{v-1}$, where $v$ is the number of variables in the factor analysis.

Upper-limit heritability was assessed simply by exploring full-sibling correlations. Doubling the correlation approximates the familiality contribution and therefore sets the upper boundary for proportion of genetic influence on a measure.

## RESULTS

### Descriptive Statistics

All tasks showed expected learning and problem-solving patterns with remarkable similarity between the two studies for comparable data. Figure 1 shows behavioral profiles and learning curves for all tasks in Study 1 and Study 2. The average number of alternations in the T-maze was 9.2, which was significantly above chance ($n$ = 245 mice who completed 14 trials within time limit; $t$-test against value of 7; $t$ = 15.9; $p < 0.0001$; see Fig. 1a). This value was 8.1 for Study 1 (using T-maze I: $t$ = 4.5, $p < 0.0001$) and 9.8 for Study 2 (using T-maze II: $t$ = 18.1, $p < 0.0001$). We attribute the difference between the two studies to differences in the mazes rather than differences in the populations as the Study 2 mice (all males) alternated more than Study 1 males ($t$ = 4.1, 2-tailed $p < 0.0001$). For the Hebb–Williams, there were significant improvements ($p < 0.05$, one-tailed paired $t$-tests) in latencies and errors (See Fig. 1b) between the first and last trials for all mazes except Study 1 batch 1 Maze 4 (latency) and Study 2 batch 1 Maze 3 (latency and errors). For the Morris hidden platform task, significant drops in latency were seen from first to last block in Study 1 batch 1 ($t = 3.3$, $p = 0.001$), Study 1 batch 2 ($t$ = 4.4, $p < 0.0001$), and Study 2 ($t$ = 10.2, $p < 0.0001$). Platform re-location (reversal trial) in Study 2 produced a significant increase in latency to the next block ($t$ = 5.2, $p < 0.0001$) followed by a learning of the new platform position as indicated by a significant decrease in latency to the next block ($t$ = 4.5, $p < 0.0001$. Similar "reversal" results obtained for Study 1 batch 2, when this task was first piloted—see Fig. 1c). For the burrowing puzzle, significant improvements ($p < 0.0001$) were seen in both Studies between trials 2 and 3 of the first burrowing puzzle (see Fig 1d; note that for Study 2 this statistic applies only to batch 1). In Study 1, there was no such one-trial learning on the second burrowing puzzle. In Study 1 batch 1, the third puzzle in

**Table III.** Consistent Sex Differences in Study 1

| Task | Performance | Mean (SD) | | Significance | |
|---|---|---|---|---|---|
| | | Males | Females | *t* | *p* |
| TM-E | M > F | −0.18 (0.82) | 0.18 (1.12) | 1.63 | 0.11 |
| BP-L | M > F | −0.23 (0.39) | 0.23 (1.32) | 2.20 | 0.03 |
| PP-L | M > F | −0.31 (0.64) | 0.31 (1.18) | 2.98 | < 0.01 |
| HW-E | M > F | −0.27 (0.94) | 0.27 (0.99) | 2.57 | 0.01 |
| HW-L | M > F | −0.46 (0.65) | 0.46 (1.07) | 4.79 | < 0.0001 |
| MH-L | M > F | −0.26 (0.94) | 0.26 (0.99) | 2.48 | 0.02 |
| First factor (*g*) | M > F | −0.50 (0.66) | 0.50 (1.04) | 5.24 | < 0.0001 |

Note that the male and female means are symmetrical about zero, this is because the scores are standardized and the male and female group sizes are equal ($n = 42$ each). TM-E = T-maze (errors). BP-L = burrowing puzzle (latency). PP-L = plug puzzle (latency). HW-E = Hebb–Williams maze (errors). HW-L = Hebb–Williams maze (latency). ML-L = Morris water maze hidden platform task (latency). All *t*-tests are two-tailed. The "First factor (*g*)" was derived from principal component factor analysis on these measures (see text).

the series showed one-trial learning ($t = 3.5$, $p < 0.0001$), but the fourth did not ($t = 0.6$, $p = 0.28$). With the plug puzzle, a significant drop in latency was seen between the first and second problem trial in Studies 1 and 2 ($t = 8.9$, $p < 0.0001$ and $t = 11.6$, $p < 0.0001$, respectively; see Fig. 1d).

In the water plus maze, there were significant decreases in latency ($t = 9.5$, $p < 0.0001$) and errors ($t = 6.1$, $p < 0.0001$) from the first to last trial (see Fig. 1e). On the object exploration task, mean exploring latency was 46.5 (s.d. = 28.0; see Fig. 1f for distribution). On the syringe task, there were highly significant increases in latency between the training and problem first problem trial in a three groups run ($t = 10.8, 7.0,$ and $9.3, p < 0.0001$), but no significant learning across problem trials (see Fig. 1g).

In Study 1, sex differences were analyzed. As shown in Table III, males significantly ($p < 0.05$) outperformed females in all tasks except spontaneous alternation. Note also that within both batches of Study 1, males outperformed females on all tasks, although not always significantly. The generation of the *g*-score shown in Table III is discussed later. The effect of albinism, which may moderate performance (Creel, 1980; Lasalle and Le Pape, 1981), was also explored in Study 1. There was seen to be no significant effect or even any overall trend between the 13 albino and 71 pigmented mice except in the Morris water maze where the albino mice underperformed ($p < 0.0001$).

## Task Reliabilities

Table IV shows trial scores in latencies and errors and also internal consistencies for the measures. Reliabilities are measured by mean correlation between trials and also Cronbach's alpha in the case of more than two trials per measure. Cronbach's alpha is a reliability coefficient based on mean inter-trial correlation and number of component trials. The range is from 0 to 1 where alpha > 0.6 is generally regarded as representing a good ratio of information to error in the whole task (composite of summed trials). Also shown are the reliability statistics re-calculated following the removal of mice that failed at any point and scored a maximum latency, as it has been suggested that non-performance of mice in behavior tasks can artificially inflate reliability statistics (Wahlsten *et al.*, 2003). The mean inter-trial correlations ranged from −0.12 to 0.55, and from 0.00 to 0.50 after exclusion of mice with one or more "failure". Corresponding Cronbach's alphas range from 0.00 to 0.96. We attribute the lower reliabilities in Study 2 primarily to the shortened tasks. Of particular note is the large failure rate of the first batch run on the syringe puzzle (mainly trial 1), hence the subsequent procedure change. Included in the Table IV are sibling correlations for the measures; these are explored further in Section "Heritablility Estimates".

## Inter-task Correlations

Table V shows the correlations among all cognitive measures with Study 1 results above the diagonal and Study 2 results below. Spontaneous alternation is coded as errors (14 minus number of alternations) and object exploration time was inversed ("additive inverse": total time minus exploration time) so that for all measures low scores indicate better performance. From the whole set of 70 mea-

**Table IV.** Reliabilities of Measures and Sibling Correlations

| | Study 1 | | | | | | | Study 2 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | All mice | | | "Failing" mice removed | | | | All mice | | | "Failing" mice removed | | |
| | Trials | *n* | *r* | Cr α | *n* | *r* | Cr α | Trials | *n* | *r* | Cr α | *n* | *r* | Cr α | Sib-corr[a] |
| T-maze (errors) | 14 | 84 | – | – | 82 | (0.22[b]) | | 14 | 169 | – | – | 163 | (0.16[b]) | 0.12 | |
| Burrowing puzzle (latency) | 11/5 | 40/44 | 0.37, 0.49 | 0.86, 0.83 | 30/42 | 0.14, 0.15 | 0.65, 0.47 | 5/1 | 78/90 | 0.29 | 0.67 | 73/88 | 0.24 | 0.62 | 0.12 |
| Plug puzzle (latency) | 2 | 84 | 0.55 | – | 78 | 0.50 | – | 2 | 168 | 0.47 | – | 158 | 0.36 | – | 0.12 |
| Hebb–Williams (latency) | 40/18 | 40/44 | 0.37, 0.31 | 0.96, 0.89 | 12/24 | 0.15, 0.14 | 0.87, 0.75 | 18/15 | 78/90 | 0.17, 0.15 | 0.78, 0.72 | 38/66 | 0.10, 0.05 | 0.67, 0.46 | 0.14 |
| Hebb–Williams (errors) | 40/18 | 40/44 | 0.12, 0.14 | 0.84, 0.74 | 12/24 | 0.14, 0.13 | 0.87, 0.72 | 18/15 | 78/90 | 0.07, 0.03 | 0.56, 0.31 | 38/66 | 0.03, 0.00 | 0.32, 0.03 | 0.04 |
| MWM learning (latency) | 32/20 | 40/44 | 0.07, 0.10 | 0.71, 0.69 | 36/40 | 0.08, 0.10 | 0.74, 0.68 | 20 | 168 | 0.05 | 0.52 | 149 | 0.04 | 0.46 | 0.12 |
| MWM reversal (latency) | | | | | | | | 7 | 168 | 0.07 | 0.33 | 165 | 0.08 | 0.36 | –0.06 |
| Water plus maze (latency) | | Not run in Study 1 | | | | | | 11 | 167 | 0.16 | 0.70 | 152 | 0.17 | 0.72 | 0.36 |
| Water plus maze (errors) | | | | | | | | 11 | 167 | 0.07 | 0.46 | 152 | 0.07 | 0.47 | 0.29 |
| Object exploration (latency) | | | | | | | | 1 | 168 | – | – | 152 | – | – | 0.24 |
| Syringe puzzle (latency) | | | | | | | | 2/3/1 | 39/39/90 | –0.12, 0.14 | 0.68 | 1/13/78 | 0.24 | 0.49 | –0.03 |

*n* = sample size, *r* = mean inter-trial correlation. Cr α = Cronbach's alpha. [a]Sibling correlations presented for batches combined in all cases, outliers removed. [b]This statistic is a Spearman–Brown corrected odd–even correlation ($=2r/[1+r]$). Where sub-batches of a study ran slightly different procedures, more than one result may be given; sample size and trial number are divided by "/" and data is separated by ",".

sures, 14 were negative and 54 were positive (three were exactly zero). All six within-arena correlations were significantly positive (with $p < 0.01$), and the remaining 15 significant correlations were positive cross-arena relationships. The mean correlation value was 0.21 for Study 1 and 0.09 for Study 2. For purely cross-arena correlations, these values were 0.18 for Study 1 and 0.06 for Study 2.

## Factor Analysis

The data shown in Table V for Study 1 and Study 2 were separately subjected to unrotated principal component factor analysis. These PCFA results are shown in Table VI. Removal of outliers over 3 standard deviations from the mean was taken as a standard necessary treatment, but both outlier-removed and untreated data are reported. For Study 1, all six measures loaded positively on the first principal component of an unrotated PCFA accounting for 35% of the variance (eigenvalue of 2.1). Removing outliers caused no notable differences; neither did repetition of the analyses with sex-regressed data (although somewhat smaller magnitudes of the first factor were obtained). Similarly for Study 2, all 11 measures loaded positively on the first principal component of an unrotated PCFA for both raw data and data with outliers removed.

However, we note that some arenas had more than one measure. This not only gives those arenas a greater representation than others, but substantial within-arena correlations can dominate the factor-analytic solution. To analyze purely cross-arena variance, another set of factor analysis was run in which each cognitive arena employed had only one representative measure. To achieve this, both measures from the puzzle box (i.e. burrowing puzzle and plug puzzle) were standardized and summed to make a puzzle box score. This was then re-standardized for ease of use. Similarly with the latencies and errors for the Hebb–Williams, the hidden task and reversal of the Morris water maze, and for the latency and error scores for the water plus maze. The spontaneous alternation, object exploration and syringe puzzle data were simply standardized. This then provided 4 standardized scores for Study 1 and 7 standardized scores for Study 2 (one score for each arena). The mean intercorrelation between these measures was 0.21 for Study 1 and 0.07 for Study 2. The principal component factor analyses for both studies are shown in Table VII. Again, all arena loadings were positive on the first factor in both studies both before and

**Table V.** Pearson's Correlations between the Diverse Set of Ability Measures

| | TM-E | BP-L | PP-L | HW-L | HW-E | MH-L | MR-L | WP-L | WP-E | OE-L |
|---|---|---|---|---|---|---|---|---|---|---|
| TM-E | | 0.10 | 0.06 | 0.22* | 0.17 | 0.14 | | | | |
| BP-L | 0.17* | | 0.52** | 0.21 | 0.12 | 0.25* | | | | |
| PP-L | 0.24** | 0.49** | | 0.30** | 0.13 | 0.05 | | Study 1 (upper) | $n=84$ | |
| HW-L | −0.05 | 0.12 | 0.04 | | 0.32** | 0.39** | | | | |
| HW-E | −0.04 | 0.00 | 0.02 | 0.37** | | 0.18 | | | | |
| MH-L | 0.00 | −0.07 | −0.07 | 0.08 | 0.18* | | | | | |
| MR-L | 0.14 | 0.17* | 0.08 | −0.14 | −0.06 | 0.26** | | Study 2 (lower) | $n=167$ | |
| WP-L | −0.08 | 0.10 | 0.06 | 0.09 | −0.01 | 0.17* | 0.21** | | | |
| WP-E | −0.09 | 0.01 | −0.05 | −0.03 | 0.05 | 0.00 | 0.06 | 0.75** | | |
| OE-L | 0.02 | 0.27** | 0.29** | 0.09 | 0.08 | 0.09 | 0.23** | 0.07 | 0.03 | |
| SP-L | 0.03 | 0.15* | 0.17* | 0.01 | −0.06 | 0.02 | 0.09 | −0.07 | −0.05 | 0.07 |

*$p<0.05$, **$p<0.01$. Study 1 above diagonal, Study 2 below. Note that values on the diagonal represent the correlations between a measure and itself (value always = 1.0), and have been removed for clarity. TM-E = T-maze (errors). BP-L = burrowing puzzle (latency). PP-L = plug puzzle (latency). HW-L = Hebb–Williams (latency). HW-E = Hebb–Williams (errors). MH-L = Morris water maze hidden platform task (latency). MR-L = Morris water maze reversal trial (latency). WP-L = water plus maze (latency). WP-E = water plus maze (errors). OE-L = object exploration (latency). SP-L = syringe puzzle (latency).

after outlier removal. The first factor accounted for 41% of the variance in Study 1 (or 36% after sex regression) and 22–23% in Study 2. Note that repeating the Study 2 analysis including only Study 1 measures (i.e. a replication) produced a first factor with all positive arena loadings and accounting for 31% of battery variance.

To test the robustness of the factor analytic structure across permutations of the battery, the factor analysis was repeated four times for Study 1 and seven times for Study 2, each time excluding a different arena. The process was then repeated with outliers removed. This resulted in 10 sets of first factor loadings (including the two original 4-measure factor analyses) for Study 1; and 16 sets (including the two original 7-measure factor analyses) for Study 2. The range of the 14 factor loadings generated for each arena in this robustness analysis is shown in Figure 2. Note that of all the battery combinations, only one (Study 2; object exploration removed, outliers included) showed a measure loading negatively on the first factor (water plus maze, −0.05). In Study

**Table VI.** First Factor Loadings for the Cognitive Measures, including Repetitions of the Analysis for Scores Corrected for Sex and Outliers

| | Study 1 | | Study 1 (sex regressed) | | Study 2 | |
|---|---|---|---|---|---|---|
| | Raw data, $n=84$ | No outliers, $n=78$ | Raw data, $n=84$ | No outliers, $n=79$ | Raw data, $n=1674$ | No outliers, $n=153$ |
| T-maze (errors) | +0.40 | +0.50 | +0.33 | +0.49 | +0.25 | +0.19 |
| Burrowing puzzle (latency) | +0.66 | +0.61 | +0.67 | +0.76 | +0.66 | +0.60 |
| Plug puzzle (latency) | +0.62 | +0.64 | +0.58 | +0.61 | +0.61 | +0.51 |
| Hebb–Williams (latency) | +0.65 | +0.73 | +0.54 | +0.63 | +0.20 | +0.14 |
| Hebb–Williams (errors) | +0.60 | +0.56 | +0.54 | +0.48 | +0.15 | +0.09 |
| MWM learning (latency) | +0.56 | +0.43 | +0.50 | +0.31 | +0.24 | +0.26 |
| MWM reversal (latency) | | | | | +0.49 | +0.46 |
| Water plus maze (latency) | | Not run in Study 1 | | | +0.53 | +0.70 |
| Water plus maze (errors) | | | | | +0.37 | +0.60 |
| Object exploration (latency) | | | | | +0.58 | +0.47 |
| Syringe puzzle (latency) | | | | | +0.23 | +0.11 |
| *Eigenvalue (%)* | 2.1 | 2.1 | 1.7 | 1.9 | 2.0 | 2.1 |
| *Percentage of variance (%)* | 35 | 34 | 29 | 32 | 18 | 19 |

Each column represents a separate principal component factor analysis. Values given are the factor loadings of the measures on the first unrotated factor. Eigenvalues and percentage of variance statistics refer to this first factor. Note that loadings may be either positive or negative and the hypothesis is that all factor loadings are positive.

**Table VII.** First Factor Loadings for the Cognitive Arenas, Including Repetitions of the Analysis for Scores Corrected for Sex and Outliers

| | Study 1 | | Study 1 (sex regressed) | | Study 2 | |
|---|---|---|---|---|---|---|
| Arena | Raw data, $n=84$ | No outliers, $n=80$ | Raw data, $n=84$ | No outliers, $n=82$ | Raw data, $n=167$ | No outliers, $n=153$ |
| T-maze | +0.51 | +0.58 | +0.49 | +0.47 | +0.37 | +0.25 |
| Puzzle box | +0.57 | +0.61 | +0.44 | +0.59 | +0.72 | +0.68 |
| Hebb–Williams maze | +0.78 | +0.81 | +0.75 | +0.74 | +0.23 | +0.23 |
| Morris water maze | +0.68 | +0.54 | +0.66 | +0.55 | +0.47 | +0.52 |
| Water plus maze | | | | | +0.13 | +0.19 |
| Object exploration | | Not run in Study 1 | | | +0.70 | +0.72 |
| Syringe puzzle | | | | | +0.37 | +0.31 |
| *Eigenvalue (%)* | 1.7 | 1.6 | 1.4 | 1.4 | 1.6 | 1.5 |
| *Percentage of variance (%)* | 41 | 41 | 36 | 36 | 23 | 22 |

Each column represents a separate principal component factor analysis. Values given are the factor loadings of the measures on the first unrotated factor. Eigenvalues and percentage of variance statistics refer to this first factor. Note that loadings may be either positive or negative and the hypothesis is that all factor loadings are positive.

1, all arenas consistently loaded highly on the first factor. In Study 2, the most consistently high-loading measures are the puzzle box (mean loading = 0.71), the Morris water maze (0.53) and the object exploration in the open field (0.71). The distribution of associated first factors for Study 1 (10 replications) had a mean eigenvalue of 1.47 and on average accounted for 46% of battery variance. The distribution of associated first factors for Study 2 (16 replications) had a mean eigenvalue of 1.48 and on average accounted for 24% of battery variance.

### Heritability Estimates

Table IV shows sibling correlations for all measures used. Outliers were removed, so sibling correlations presented are for between 79 and 83 pairs. The mean sibling correlation was 0.13. Although this is not statistically significant, it should be noted that the 11 measures showed a standard deviation of 0.13 about this value. This tight distribution of low but positive values is significantly above zero ($t=3.5$, $p<0.005$, one-tailed). Doubling a sibling correlation provides an upper-limit heritability estimate. It is "upper-limit" because shared maternal/litter effects cannot be discounted. The mean upper-limit heritability estimate for these individual measures is therefore 26%. However, note that the low reliabilities of these tasks set a lowered ceiling for these correlations and subsequent heritability estimates.

Sibling correlations were also calculated for the *g*-scores derived from all the principal component factor analyses on Study 2 described above. Sibling correlations for these 16 different *g*-scores had a mean of 0.17 (s.d. = 0.10). Recalculating the sibling correlations with outliers (average 1.75 outliers per *g*-score) removed yielded a mean sibling correlation of 0.21. Doubling these sibling correlations produces an upper-limit heritability estimate in the range of 34–42%.

An alternative method to explore the consistency of the factor structure and heritability of the tentative *g* factor is to assess the correlation matrix and factor structure in two populations—with each sibling pair having one member in each population. This is shown in Figure 3. This results in a predominantly positive matrix for the "sib 1" population of 83 mice, and a similar correlation matrix for the population of their 84 co-sibs, denoted "sib 2". Principal component factor analysis on these two groups produces similar results to before, although it is noted that there was one negative loading (syringe puzzle) in the group "sib 2". Nevertheless, both sets of data produced first factors accounting for 23–24% of the variance, and the sib 1 scores correlated 0.17 with the sib 2 scores independently generated. Again, to check the robustness of the pattern, this was repeated with outliers (over 3SD on any task) removed. Results obtained were similar with *g*-factors accounting for 22% and 25%, and correlating 0.20 (yet it is noted that now the T-maze produced a slightly negative factor loading for the "sib 1" group).

## DISCUSSION

The results reported here show a factor structure indicative of the presence of a *g* factor and a modest degree of familiality for this factor and for the cognitive tasks employed. The large dataset generated by

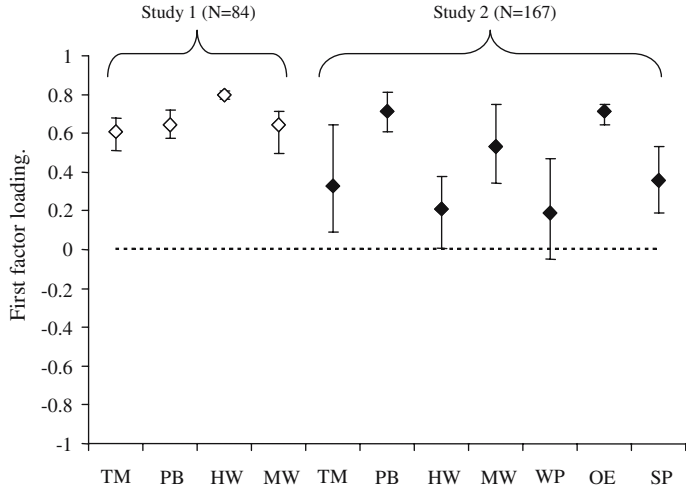


**Fig. 2.** Robustness analysis of factor loading structure showing the range of factor loadings for each arena. Data points and bars represent means and full ranges of factor loadings for each arena. These were generated by the full factor analysis, and then replications each with the exclusion of a single arena. This was then repeated with outliers removed. There were therefore 8 replications for the factor analysis in Study 1 and 14 replications for the factor analysis in Study 2. All loading distributions are significantly departed from zero (see text). TM = T-maze, PB = puzzle box, HW = Hebb–Williams maze, MW = Morris water maze, WP = water plus maze, OE = object exploration, SP = syringe puzzle.

the study was explored in several ways; with replications using different populations, different analysis methods, different data treatments, and different battery compositions. Nevertheless, results were seen to be consistent across all these permutations. Low internal reliabilities of the measures limited the size of the tentative *g*-factor to 36% in the first study (following sex-regression) and 23% in the replication (or 31% if only the same tasks as in the first study are analyzed). Sex differences were seen in general cognitive task performance with males outperforming females on all tasks, and this served to inflate the magnitude of the *g*-factor in Study 1.

A novel feature of the study is the use of siblings. Despite the low internal reliabilities of the tasks employed, the mean sibling correlation for the cognitive measures was 0.13, with the *g*-factor sibling correlations ranging from 0.17 to 0.21. Doubling this difference gives a familiality estimate of around 26% for the component tasks and 34–42% for the derived *g*-factors. This is also the "upper-limit of heritability" estimate important for consideration when cognitive QTLs are investigated in this outbred population of mice run on this battery. Note that if these estimates were expressed as a proportion of reliable variance then they would be substantial. However, it makes more sense practically to improve reliabilities rather than adjust estimates. It had been planned that genetic and maternal-environmental factors could also be separated by use of the half-sibling design in the second study. However, given the sample size and the relatively low magnitude of full-sibling correlations, this aspect could not be productively explored with these data. Even so, opportunities for such designs in combination with more reliable physiological and behavioral measures are clear.

Although the results of the Study 1 agree with previous findings showing mean cross-task correlations in the region of 0.20 and supporting the notion of a general cognitive ability across different motivations (Galsworthy *et al.*, 2002; Matzel *et al.*, 2003), the weakness of some inter-correlations and loadings in Study 2 also shows these studies to be compatible with other recent results showing non-uniformly aligned factor loadings (Locurto *et al.*, 2003). In fact, the mean cross-arena correlation reported here for Study 2 is the lowest yet published and we attribute this partly to the shortening of many tasks. Yet as with our previous report (Galsworthy *et al.*, 2002), the statistical summaries showed large preponderance of support for the *g* hypothesis that was also backed up by individual performances. In Study 2, for example, one mouse (mouse number 12) ranked 3rd in the puzzle box, 2nd in the Morris water maze and 5th in the object exploration—out of 167 mice.

The size of the study renders many specific aspects and results to discuss. Sex differences will not be

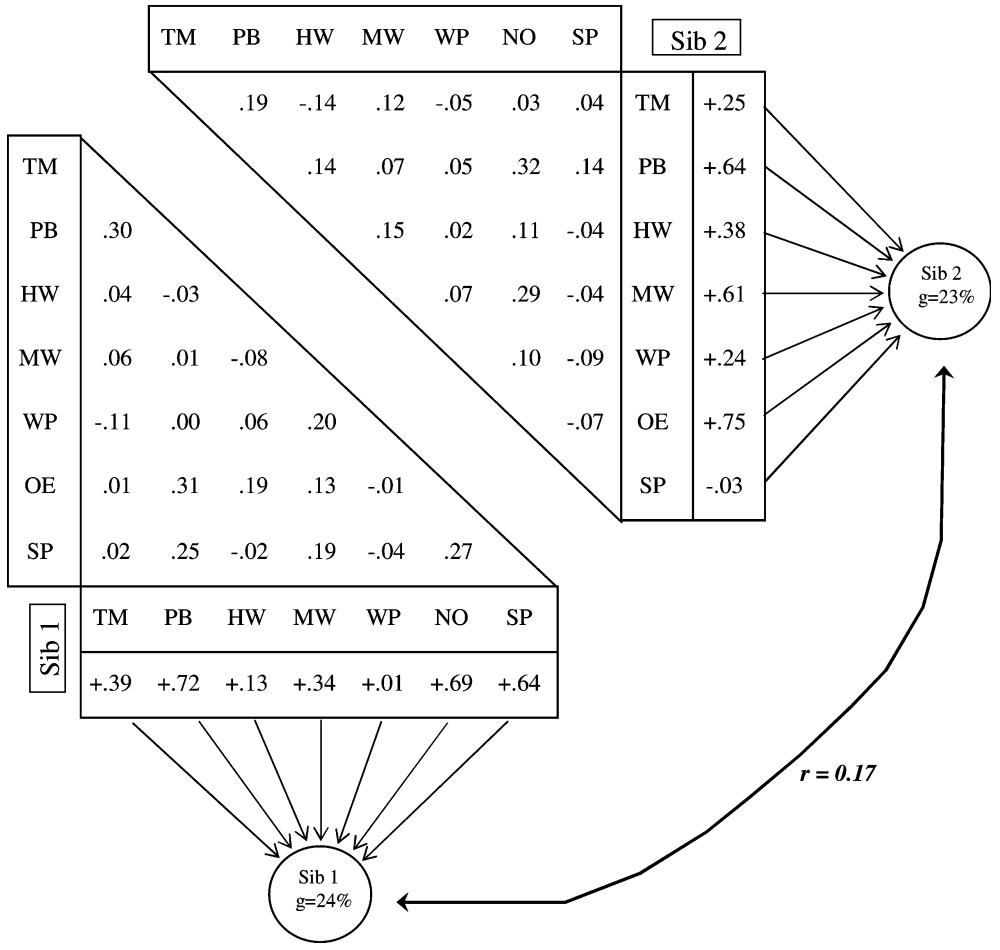


**Fig. 3.** Correlation matrices and consequential "*g*" factor loadings with sibling pairs split into independent groups. $n = 83$ for "Sib 1"—a group of unrelated mice; $n = 84$ for "Sib 2"—their siblings. $n = 82$ pairs for the sibling correlation ($r = 0.17$) given between the two "*g*" factors extracted by unrotated principal component factor analysis (Long boxes list loadings for the above/adjacent arenas—circles are principal components with percentage of variance accounted for shown). TM = T-maze, PB = puzzle box, HW = Hebb–Williams maze, MW = Morris water maze, WP = water plus maze, OE = object exploration, SP = syringe puzzle.

discussed here except to mention that other populations of HS mice have shown better performance from females (Locurto *et al.*, 2003). More uniquely, as one of the few recent studies in mouse psychometrics, we hope the research reported here opens more awareness of choice of tasks, choice of methodology for data analysis and importance of quantitative information. Dealing first with the choice of tasks; most "cognitive" tasks here were chosen because they appeared to form a balanced set which was inexpensive to build and run—and diverse in measurement. This allowed for a larger battery which could cover various cognitive, sensory, motor and motivational angles. Nevertheless, the tasks consumed many experimenter hours and so were shortened for Study 2, where it appeared appropriate, in order to accommodate the larger number of subjects. This appeared to affect the reliabilities and usefulness of some tasks more than others. For example, the Hebb–Williams maze was much shortened from Study 1 to Study 2 and consequently both reliability and factor loading suffered. By comparison, the burrowing puzzle in the puzzle box was also much shortened, but still held up well in the second study. In Study 2, the object exploration task also appeared to be very central to the "general" factor derived. This task was quick, easy to run and included within a "cognitive" battery as object exploration or curiosity has been nominated as a correlate of general cognitive performance in mice (Matzel *et al.*, 2003),

rats (Anderson, 1993) and human infants (Bornstein and Sigman, 1986). This was corroborated in the present study, but of course raises the awkward question of whether it is the exploratory curiosity predominantly driving good task performance, or whether superior cognitive functioning manifests itself partly in greater inquisitiveness.

As recent attention turns to the use of batteries and their standardization for behavioral phenotyping of mice (e.g., Brown *et al*., 2000; Wahlsten, 2001), there is a need for such psychometric research to show that the tasks in any such battery are reliable, heritable and share meaningful covariance. Two tasks from the current set that seem to show this property are the burrowing and plug puzzles within the puzzle box. Across both studies, puzzle box tasks showed a centrality in the cognitive battery. Other quick, low-stress and naturalistic tasks explored by other researchers may also show such properties. Examples include the olfactory learning task reported by Matzel *et al*. (2003), the "detour task" used by Locurto *et al*. (2003), maze running with home cages as a goal boxes reported by Blizard *et al*. (2003), or even automated learning paradigms within the home cage itself (Galsworthy *et al*., 2005). Cognitive arenas which are designed to reduce emotional variance and to employ motivations and demands which are more species-salient are relatively new to mouse cognitive assessment and should be psychometrically explored alongside more traditional tasks.

The data reported here cannot be taken as final proof of a general cognitive ability that will influence any task which has a cognitive element. There is a long way to go before understandings of the architecture of mouse cognition can rival the data-driven hierarchical model of human cognitive variation widely accepted today (e.g. Gustaffson, 1984; Carroll 1993). Much more exploration is needed, and failures to produce positive manifolds or uniform-direction factor loadings do not necessarily prove that general cognitive influences are not present in the data. The measures might only weakly tap cognitive influences and be swamped by other influences; a particularly destructive circumstance would be where a non-cognitive trait serves to assist good performance in one task and impair good performance in another, thus contributing a stable negative correlation between the tasks to counter any positive cognitive correlation that may exist between them. Therefore, a zero correlation between two "cognitive" tasks does not equate to zero correlation between their cognitive elements. For this reason, it should be stated that "no *g*" is not an adequate default conclusion in the face of data which gives *g* no support. Proving the "no *g*" case is an equally difficult endeavor: at least two cognitive factors would have to be independently validated in relevant batteries spanning different motivations and sensory demands before those two factors were shown to be uncorrelated. That would then actively evidence independent cognitive processes. We note in this dataset (Study 2; $n = 167$) that a first factor derived from the water tasks correlated 0.12 with a factor derived from the land tasks, again evidencing some commonality of measurement.

Previous studies of cognitive tasks indicate little or no association of general cognitive performance factors with anxiety (Galsworthy *et al*., 2002) or activity (Locurto and Scanlon, 1998; Galsworthy, 2003; Locurto *et al*., 2003; Matzel *et al*., 2003) indices. However, this does not rule out the influence of non-cognitive factors within the measures. Furthermore, low inter-arena correlations were not merely due to differences in motivation employed as even tasks based on very similar principles (e.g. Morris water maze and water plus maze) are seen to have low correlations between them. Perhaps it is the case that many tasks are not only influenced by confounding traits, but are also unreliable due to the stress producing unpredictable responses. The variety of strategies available for task solution in some tests may also produce large task-specific variance.

Whilst mean sibling correlations for individual tasks were variable, the factor extracted from the battery showed a higher mean sibling correlation. This indicates that batteries may provide more reliable scores than individual tasks, not only on face value, but also for purposes of exploring genetic and environmental origins of behavior traits. As Locurto *et al*. (2003) note, human *g* batteries developed largely by keeping reliable tasks that increased battery diversity but still correlated well with other tasks. By keeping those tasks that show good face validity, good psychometric properties and good heritability, it is hoped that the mouse will become a more powerful model within which to explore the functional genomics of human cognitive abilities and disabilities (Plomin, 2001).

In summary, there is now accumulating evidence for a general cognitive ability factor—or at least a general cognitive task performance factor in mice. This study has also opened a quantitative behavioral genetics angle to complement the new psychometrics. However, it is becoming clear that too many currently used tasks show weak learning, low reliabilities

and with results varying between labs (Crabbe *et al*., 1999; Wahlsten, 2001). We conclude that tasks are needed which are psychometrically validated and heritable. Developing better cognitive tasks will certainly facilitate understanding of mouse behavior *per se* and also open the door to more refined explanations of the effects of genetic, environmental or experimental variables on behavior. The research reported here indicates that a powerful individual differences animal model to aid research into human cognitive abilities and disabilities is attainable.

## APPENDIX. RE-ANALYSIS OF BAGG (1920) DATA.

Bagg's 93 yellow and white mice from various families were run through a series of cognitive tests. Although only one cross-measure correlation was calculated, the individual mean scores for some 80 animals were presented in the paper. A re-analysis of this data for the 71 reported animals that had full sets of data is shown here:

Mice were first run through a "maze test" (MT) which comprised two sections in series, each with two doors that could be pushed open. The correct choice was door A (section 1) then door B (section 2), which would allow mice to escape into a community area with bedding and food. An "interference test" (IT) was then run in the same arena whereby mice had to now go through door B in section 1 then door A in section 2—the other two doors being locked. The mice were then tested in a different "multiple choice" (MC) arena where they were presented with five doors. One door led to a community area similar to before, but the other four locked doors were also punished with a mild foot shock. Finally, mice were returned to the original maze for a "retention test" (RT). Latency (L) and error (E) measures were taken for all four tasks. Presented below are the intercorrelations of these measures and subsequent results of a principal component factor analysis (unrotated):

## ACKNOWLEDGEMENTS

The authors would like to thank Hans-Peter Lipp, David Wolfer and Lou Matzel for their comments on previous versions of this manuscript. We also thank Jerry Salazar at IBG for tailoring animal production to our needs. M. Galsworthy is supported by the Swiss National Foundation and the NCCR "Neural Plasticity and Repair". The work was funded by US NIH grant HD27694 to R. Plomin and UK grant MRC G0000170 to L. Schalkwyk.

## REFERENCES

Anderson, B. (1993). Evidence from the rat for a general factor that underlies cognitive performance and that relates to brain size: intelligence? *Neurosci. Lett*. **153**:98–102.

Ashe, K. H. (2001). Learning and memory in transgenic mice modeling Alzheimer's disease. *Learning Memory* **8**(6), 301–308.

Bagg, H. J. (1920). Individual differences and family resemblances in animal behavior. *Arch. Psychol*. **43**:1–58.

Blizard, D. A., Klein, L. C., Cohen, R., and McClearn, G. E. (2003). A novel mouse-friendly cognitive task suitable for use in aging studies. *Behav. Genet*. **33**(2), 181–189.

Bornstein, M. H., and Sigman, M. D. (1986). Continuity in mental-development from infancy. *Child Develop*. **57**(2), 251–274.

Brody, N. (1992). *Intelligence* (2nd ed.). New York, NY: Academic Press.

Brown, R. E., Stanford, L., and Schellinck, H. (2000). Mouse IQ: developing standardized behavioral tests for knockout and inbred mice. *ILAR J*. **41**:163–174.

Carroll, J. B. (1993). *Human Cognitive Abilities*. New York: Cambridge University Press.

Cohen, J. (1988). *Statistical Power for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.

Crabbe, J. C., Wahlsten, D., and Dudeck, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science* **284**(5420), 1670–2.

Creel, D. (1980). Inappropriate use of albino animals as models in research. *Pharmacol., Biochem. Behav*. **12**:969–977.

**Appendix Table**

| | MT-L | MT-E | IT-L | IT-E | MC-L | MC-E | RT-L | PCFA* | Loading |
|---|---|---|---|---|---|---|---|---|---|
| MT-L | – | | | | | | | MT-L | +0.75 |
| MT-E | 0.87 | – | | | | | | MT-E | +0.81 |
| IT-L | 0.64 | 0.71 | – | | | | | IT-L | +0.89 |
| IT-E | 0.56 | 0.62 | 0.90 | – | | | | IT-E | +0.83 |
| MC-L | 0.21 | 0.22 | 0.33 | 0.27 | – | | | MC-L | +0.86 |
| MC-E | 0.35 | 0.40 | 0.54 | 0.47 | 0.81 | – | | MC-E | +0.81 |
| RT-L | 0.52 | 0.59 | 0.73 | 0.61 | 0.47 | 0.61 | – | RT-L | +0.56 |
| RT-E | 0.48 | 0.58 | 0.60 | 0.62 | 0.42 | 0.51 | 0.79 | RT-E | +0.72 |

*Principal component factor analysis (PCFA) applied to the data yields a first factor of eigenvalue 4.91 accounting for 61% of the variance in this set of measures. All loadings on the first factor are positive.

Crinella, F. M., and Yu, J. (1995). Brain mechanisms in problem solving and intelligence: a replication and extension. *Intelligence* **21**(2), 225–246.

Deary, I. J. (2000). *Looking Down on Human Intelligence: From Psychometrics to the Brain*. New York: Oxford University Press.

D'Hooge, R., and De Deyn, P. P. (2001). Applications of the Morris water maze in the study of learning and memory. *Brain Res. Rev.* **36**(1), 60–90.

Galsworthy, M. J., Paya-Cano, J. L., Monleón, S., and Plomin, R. (2002). Evidence for general cognitive ability (g) in heterogeneous stock (HS) mice and an analysis of potential confounds. *Genes, Brain Behav.* **1**(2), 88–95.

Galsworthy, M. J. (2003). A psychometric and quantitative genetic study of cognitive task performance in a heterogeneous stock (HS) population of *Mus musculus*. Unpublished Ph.D. thesis. University of London, UK.

Galsworthy, M. J., Amrein, I., Kuptsov, P., Polataeva, I., Zinn, P., Rau, A., Vyssotski, A., and Lipp, H.-P. (2005). A comparison of wild-caught wood mice and bank voles in the Intellicage: assessing exploration, daily activity patterns and place learning paradigms. *Behav. Brain Res.* **157**(2), 211–217.

Gerlai, R. (1998). A new continuous alternation task in T-maze detects hippocampal dysfunction in mice. A strain comparison and lesion study. *Behav. Brain Res.* **95**:91–101.

Gustaffson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence* **8**:179–203.

Hebb, D. O., and Williams, K. (1946). A method of rating animal intelligence. *J. Genet. Psychol.* **34**:59–65.

Lassalle, J.-M., and Le Pape, G. (1981). Differential effects of the albino gene on behavior according to task, level of inbreeding, and genetic background. *J. Comp. Physiol. Psychol.* **95**:655–662.

Lipp, H.-P., and Wolfer, D. P. (1998). Genetically modified mice and cognition. *Curr. Opin. Neurobiol.* **8**:272–280.

Locurto, C., and Scanlon, C. (1998). Individual differences and a spatial learning factor in two strains of mice (*Mus musculus*). *J. Comp. Psychol.* **112**:344–352.

Locurto, C., Fortin, E., and Sullivan, R. (2003). The structure of individual differences in Heterogeneous Stock mice across problem types and motivational systems. *Genes, Brain Behav.* **2**(1), 40–55.

Mackintosh, N. J. (1998). *IQ and Human Intelligence*. Oxford: Oxford University Press.

Matzel, L. D., Han, Y. R., Grossman, H., Karnik, M. S., Patel, D., Scott, N., Specht, S. M., and Gandhi, C. C. (2003). Individual differences in the expression of a 'general' learning ability in mice. *J. Neurosci.* **23**(16), 6423–6433.

McClearn, G. E., Wilson, J. R., and Meredith, W. (1970). The use of isogenic and heterogenic mouse stocks in behavioral research. In G. Lindzey and D. Thiessen (eds.), *Contributions to Behavior Genetic Analysis: The Mouse as a Prototype*. New York: Appleton Century Crofts, pp. 3–22.

Meunier, M., Saint-Marc, M., and Destrade, C. (1986). The Hebb–Williams test to assess recovery of learning after limbic lesions in mice. *Physiol. Behav.* **37**(6), 909–913.

Misslin, R., and Ropartz, P. (1981). Responses in mice to a novel object. *Behaviour* **78**:169–177.

Morris, R. G. M. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *J. Neurosci. Meth.* **11**:46–60.

Plomin, R. (1999). Genetic research on general cognitive ability as a model for mild mental retardation. *Int. Rev. Psychiat.* **11**:34–36.

Plomin, R., DeFries, J. C., McClearn, G. E., and McGuffin, P. (2001). *Behavioral Genetics* (4th ed.). New York: Worth Publishers.

Plomin, R. (2001). The genetics of *g* in human and mouse. *Nat. Rev. Neurosci.* **2**:136–141.

Rabinovitch, M. S., and Rosvold, H. E. (1951). A closed-field intelligence test for rats. *Can. J. Psychol.* **5**:122–128.

Rencher, A. C. (1995). *Methods of Multivariate Analysis*. New York: John Wiley & Sons.

Spearman, C. (1904). 'General intelligence' objectively determined and measured. *Am. J. Psychol.* **15**:201–293.

Thompson, R., Huestis, P. W., Bjelajac, V. M., Crinella, F. M., and Yu, J. (1989). Working memory in young rats with lesions to the "general learning system". *Psychobiology* **17**:285–292.

Thompson, R., Huestis, P. W., Shea, C. N., Crinella, F. M., and Yu, J. (1990). Brain structures important for solving a sawdust-digging problem in the rat. *Physiol. Behav.* **48**:107–111.

van Gaalen, M. M., and Steckler, T. (2000). Behavioural analysis of four mouse strains in an anxiety test battery. *Behav. Brain Res.* **115**(1), 95–106.

Wahlsten, D. (2001). Standardizing tests of mouse behavior: reasons, recommendations, and reality. *Physiol. Behav.* **73**:695–704.

Wahlsten, D., Rustay, N. R., Metten, P., and Crabbe, J. C. (2003). In search of a better mouse test. *Trends Neurosci.* **26**(3), 132–136.

Edited by Marty Hahn